인지과학, 제24권 제2호
Korean Journal of Cognitive Science
2013, Vol. 24, No. 2, 95～109.

좌최장일치법과 HMM을 결합한 경량화된 한국어 형태소 분석*

강 상 우† 양 재 철 서 정 연
서강대학교 삼성전자 서강대학교
컴퓨터학과 소프트웨어 센터 컴퓨터학과

본 논문에서는 제한된 자원을 사용하는 기기에 적합한 경량화된 한국어 형태소 분석 및 품사 부착 방법을 제안한다. 관련된 초기 연구로는 규칙에 기반을 둔 방법들이 적용되었으나 최근에는 통계에 기반을 둔 방법들을 중심으로 연구되고 있다. 계산 처리 능력과 사용 가능한 메모리가 제한되는 환경에서는 규칙에 기반을 둔 방법보다 상대적으로 많은 자원을 사용하는 통계에 기반을 둔 방법을 사용하여 형태소 분석 및 품사 부착을 수행하기에는 한계가 있다. 본 논문에서는 기존의 규칙에 기반을 둔 형태소 분석 방법인 좌최장일치법을 개선하여 형태소 분석을 수행하고, 통계적인 방법인 hidden Markov model을 축소하여 형태소 품사 부착을 수행한다. 제안하는 방법은 기존의 hidden Markov model을 사용한 시스템과 유사한 성능을 보여주며 소량의 메모리 사용과 월등히 빠른 속도로 형태소 분석 및 품사 부착을 수행할 수 있다.

주제어 : 형태소 분석, 품사 부착, 좌최장일치법, HMM, 모바일 기기

---

* 본 연구는 지식경제부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음[10041678, 다중영역 정보서비스를 위한 대화형 개인 비서 소프트웨어 원천 기술 개발].

† 교신저자: 강상우, 서강대학교 컴퓨터학과, 연구분야: 자연어처리
E-mail: swkang@sogang,.ac.kr

서 론

최근 휴대폰, PMP 등의 다양한 개인화된 소형 기기들의 활용도가 높아지고 자연어처리의 응용 분야가 확대되고 있다. 하지만 이러한 기기들은 계산 능력과 사용 가능한 메모리의 부족 등 많은 제약 사항이 따른다. 따라서 이러한 제한된 자원을 갖는 환경에서 자연어 처리를 수행하기 위해서는 제한된 자원에서 효과적인 처리 방법이 필요하다. 자연어 처리의 여러 요소 기술 중 형태소 분석 및 품사 부착은 다양한 자연어 처리의 응용 분야에 중요한 기반 기술로 사용되기 때문에 제한된 자원을 사용하는 환경에서 효과적으로 형태소 분석 및 품사 부착을 수행하는 방법이 필요하다. 형태소 분석이란 주어진 입력문장 또는 어절을 최소 의미 단위인 형태소로 분리하는 작업이다. 이러한 형태소 분석 결과에서 가장 적합한 형태소의 조합과 품사 정보를 선택하는 작업을 품사 부착이라 한다. 형태소 분석 및 품사 부착의 결과는 정보 검색, 정보 추출, 기계 번역 등 자연어 처리의 여러 응용 분야에서 중요하게 사용된다.

형태소 분석은 80년대부터 지속적으로 이루어져 왔으며 연구 초기에는 주로 규칙에 기반을 둔 형태소 분석에 대한 연구들이 진행되었다. 규칙에 기반을 둔 형태소 분석 연구들은 수작업으로 획득한 규칙을 이용하여 형태소 분석을 수행하였다[1-3]. 하지만 형태소 분석에 적용되는 모든 가능한 규칙을 획득하기 어렵고, 규칙 획득에 큰 비용이 드는 단점이 있다. 이러한 단점을 극복하기 위하여 최근 연구들은 통계에 기반을 둔 방법을 통하여 형태소 분석을 접근하였다. 통계에 기반을 둔 방법은 대량의 말뭉치로부터 추출한 통계 정보를 이용하여 자동으로 형태소 분석에 적용되는 규칙을 획득하고, 이를 기반으로 형태소 분석을 수행한다[4-7].

품사 부착에 대한 연구들 또한 초기에는 규칙에 기반을 둔 방법들을 적용하였다[8,9]. 이러한 방법들은 초기 형태소 분석 연구와 마찬가지로 중의성 해소에 필요한 모든 규칙을 획득하기 어렵고 많은 비용이 소모된다. 품사 부착의 통계적 접근은 형태소 분석과 유사한 장점을 가지며 기계 학습을 통하여 말뭉치로부터 통계 정보를 획득하고 획득한 정보를 기반으로 품사 부착을 수행한다. hidden Markov model(HMM)은 품사 부착에 사용되는 대표적인 통계에 기반을 둔 모델이다[10,11]. HMM을 이용한 방법은 품사 부착 정확도가 높다는 장점을 갖지만, 높은 복잡도와

다량의 통계 정보를 필요로 하기 때문에 규칙에 기반을 둔 방법과 비교하여 속도가 느린 단점이 있다.

본 논문에서는 규칙에 기반을 둔 모델과 통계에 기반을 둔 모델의 장점을 결합한 형태소 분석과 품사 부착 방법을 제안한다. 제안하는 방법은 기존의 규칙에 기반을 둔 형태소 분석 방법인 좌최장일치법을 개선하여 형태소 분석을 수행하고 통계적인 방법인 HMM을 축소 적용하여 형태소 품사 부착을 수행한다. 제안된 방법은 기본적으로 규칙에 기반을 둔 형태소 분석을 사용하기 때문에 분석 속도가 빠르고, 축소된 HMM을 결합함으로써 품사 분석 성능의 저하를 최소화할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 한국어 형태소 분석 및 품사 부착에 관련된 기존 연구에 대해 살펴본다. 3장에서는 개선된 좌최장일치법을 이용한 형태소 분석 방법에 대해서 설명하고 4장에서는 3장에서 제안한 개선된 좌최장일치법에 의하여 생성된 형태소 분석 후보 열을 축소된 HMM 기반 모델에 적용한 품사 부착 방법을 설명한다. 5장과 6장에서는 각각 실험 및 결과와 결론을 제시한다.

관련 연구

한국어 형태소 분석은 입력된 문장으로부터 모든 가능한 형태소 분석 결과를 생성하는 과정이다. 한국어 형태소 분석에 대한 연구는 80년대부터 많은 연구가 이루어져 왔으며 연구 초기에는 형태소 분석 알고리즘에 관한 것이 주를 이루었다.

김성용은 접속정보를 이용한 Tabular Parsing 방법을 제안하였다[12]. 이는 Triangular Table을 이용한 Parsing 방법으로 어절에서 가능한 모든 형태소 분석 결과를 만드는 방법이다. 이 방법은 모든 가능한 형태소 분석 결과를 만들 수 있는 장점이 있지만, 모든 가능한 형태소 분석 결과를 생성하기 때문에 중의적인 분석 결과를 갖는 어절을 분석하는데 많이 시간이 소모되고 형태소 분석 결과 저장에 필요한 메모리를 많이 요구한다는 단점을 가진다. 최장일치법은 입력 어절에 대한 첫 번째 형태소 분석 결과를 생성하는 방법이다[13, 14]. 좌최장일치법은 어절의 좌측부터 부분 형태소 열 중에서 가장 긴 형태소를 우선적으로 선택하여 형태소

간의 접속이 가능한지 여부를 검사하는 방법이다. 이 방법은 하나의 형태소 분석 결과만을 생성하기 때문에 다른 형태소 분석 방법과 비교하여 속도가 월등히 빠르다는 장점이 있지만, 동일 형태의 어절에 대하여 항상 하나의 결과만을 제공한다는 단점을 가진다. 예로, 어절 “말한”을 좌최장일치법을 이용하여 형태소 분석한 결과는 “말/일반명사+한/일반명사”이지만 올바른 분석결과는 “말/일반명사+하/동사파생접미사+ㄴ/관형형전성어미”이다. 올바른 분석 결과인 “하/동사파생접미사+ㄴ/관형형전성어미” 대신 가장 긴 형태소인 “한/일반명사”가 선택되기 때문이다. 이러한 단점을 보완하기 위해서 기분석 어절 사전 구축과 유지보수의 문제를 해결하기 위해 세종형태의미 말뭉치로부터 기분석 부분 어절사전을 구축하는 방법이 제안되었다[15].

한국어 형태소 품사 부착은 한국어 형태소 분석 과정에서 생성된 형태소 분석 결과들 중에서 최적의 형태소 분석 결과를 선택하는 과정이다. 형태소 품사 부착 방법은 크게 규칙에 기반을 둔 방법과 통계에 기반을 둔 방법으로 나눌 수 있다. 규칙에 기반을 둔 방법은 수작업으로 획득한 결정적 규칙의 집합을 사용하는 방법이다. 이 방법은 규칙 획득에 많은 비용이 들고, 중의성 해소에 필요한 모든 규칙을 얻기 어렵다는 단점을 가진다. 통계에 기반을 둔 방법은 말뭉치로부터 추출한 통계 정보를 이용하는 방법이다. 학습 말뭉치에서 기계학습을 통하여 통계 정보를 추출하고, 이를 바탕으로 형태소 품사 부착을 수행한다. 최근에는 컴퓨터의 성능이 급속도로 증가하여, 대량의 말뭉치를 사용한 통계에 기반을 둔 접근법이 주를 이루게 되었다. 김재훈 외[10]은 한국어에 HMM을 적용하여 한국어의 특성에 의해 발생하는 입력 열이 여러 가지로 발생하는 문제를 해결하였고 신상현 외[16]은 통계와 규칙에 기반을 둔 2단계 형태소 품사 부착 방법을 제안하였다. 통계적인 방법으로 해결되지 않는 오류들을 자동 생성된 규칙들을 사용하여 해결하였다. 또한 접속 정보를 이용하지 않고 자동으로 학습한 통계정보만을 이용하여 형태소 분석 및 품사 부착 방법이 제안되었다[17]. 이재성은 형태소의 분리와 복원을 동시에 수행하는 모델과 품사 부착 모델을 사용하는 기존의 2단계 분석 모델을 형태소 분리, 형태소 복원, 품사 부착 3개로 나누어 복잡도를 줄이는 방법을 제안하였다[18].

시스템 자원이 제한되고 계산 능력이 상대적으로 부족한 환경에서는 형태소 분석 및 품사 부착에 대한 연구가 많이 이루어지지 않았다. 다만 형태소 분석과 유

사한 방법이 많이 사용되는 자동 띄어쓰기의 경량화에 대한 연구가 있었으며 통계 정보와 오류 교정 규칙을 2단계로 적용하는 하이브리드 방법을 연구되었다[19]. 김학수는 저 사양 단말기기에 적합한 패턴 매칭 기반의 자동 띄어쓰기 방법을 제안하였다[20]. 이 방법은 음절 *n*-그램 사전을 사용하여 복잡한 확률계산 없이 단순 검색을 통해서 자동 띄어쓰기를 수행하였다. 본 논문에서는 제한된 자원을 갖는 기기에서 형태소 분석 및 품사 부착을 위한 새로운 방법을 제안한다.

개선된 좌최장일치법을 이용한 형태소 분석

규칙에 기반을 둔 모델인 좌최장일치법[14]은 빠른 분석이 가능하지만 동일 형태의 어절에 대하여 항상 하나의 결과만을 제공하는 단점을 가진다. 그림 1의 예는 좌최장일치법에 의하여 발생한 오류의 예를 보여준다. 어절 “필요한”에 대하여 좌최장일치법의 분석 결과는 “필요/일반명사+한/일반명사”로 한 개만을 제시한다. 하지만 올바른 형태소 분석 결과는 “필요/일반명사+하/동사파생접미사+ㄴ/관형형전성어미”이다. 이러한 현상은 좌최장일치법의 특성에 의한 결과로, 올바른 분석 결과인 “하/동사파생접미사+ㄴ/관형형전성어미”가 최장 형태소인 “한/일반명사” 보다 순위가 낮음으로써 발생한다.

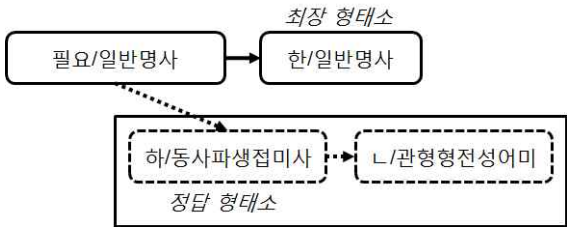


그림 1. 좌최장일치법에 의한 오류의 예

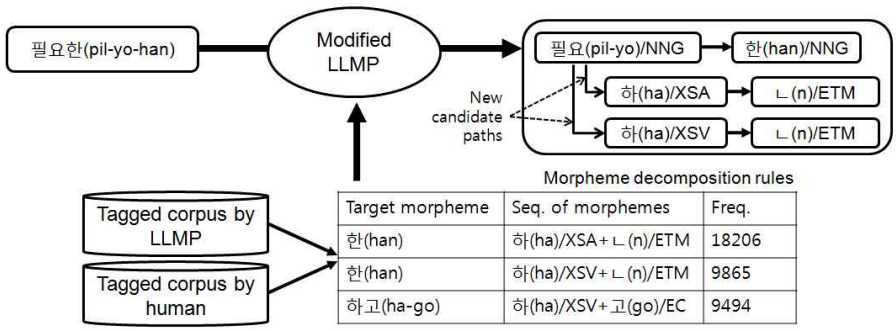
그림 1과 같이 정답 형태소 열이 잘못된 최장 형태소보다 순위가 낮음으로써 선택되지 않는 오류들이 발생한다. 이러한 문제를 해결하기 위해서는 정답을 포함할 수 있는 형태소 열 후보들을 생성해야 한다. 이를 위해 준말 처리, 불규칙 용언

표 1. 부분 형태소 열 사전의 예

부분 문자열	부분 형태소 열	발생 빈도	누적 빈도
한	하/XSA(형용사파생접미사)+ㄴ/ETM(관형형전성어미)	18206(5.56%)	5.56%
한	하/XSV(동사파생접미사)+ㄴ/ETM(관형형전성어미)	9865(3.01%)	8.57%
하고	하/XSA(형용사파생접미사)+고/EC(연결어미)	9494(2.90%)	11.46%
할	하/XSV(동사파생접미사)+ㄹ/ETM(관형형전성어미)	8816(2.69%)	14.16%
적인	적/XSN(명사파생접미사)+이/VCP(긍정지정사)+ㄴ/ETM(관형형전성어미)	7615(2.32%)	16.48%
들이	들/XSN(명사파생접미사)+이/JKS(주격조사)	7085(2.16%)	18.64%
...			

의 원형 복원 등 추가적인 작업이 필요하다[14]. 이러한 방법들은 자원 사용을 증가시키기 때문에 제한된 자원을 제공하는 환경에 적용하기 어렵다. 본 논문에서는 이러한 문제를 해결하기 위하여 *부분 형태소 열 사전*을 사용한다. *부분 형태소 열 사전*은 동일한 형태소 분리가 일어나는 문자열에 최장일치법에 의해 선택된 한 개의 후보 이외에 정답으로 선택될 수 있는 추가적인 후보들을 제공한다. *부분 형태소 열 사전*은 최장 형태소에 의하여 올바른 분석 결과의 순위가 낮아지는 오류들을 수집하여 이런 오류들에서 나타나는 정답 부분 형태소 열을 수집하여 구성된다. *부분 형태소 열 사전*에 포함된 부분 형태소 열이 어절에 존재 할 때, 부분 형태소 열의 형태소 분석 결과를 추가해주는 역할을 한다. *부분 형태소 열 사전*은 표 1과 같은 구조를 가진다.

좌최장일치법에 의하여 생성되는 모든 오류에 대하여 *부분 형태소 열 사전*을 생성하는 것은 제한된 환경에서 효율적으로 동작하기 어렵다. 낮은 빈도의 정답 부분 형태소 열을 위해 모든 후보 형태소 열을 사전으로 구축하게 된다면 좌최장일치법의 장점인 속도 향상을 기대할 수 없다. 또한 품사 부착 시에 과도하게 생성된 부분 형태소 열들로부터 정답 열을 선택할 수 있는 확률은 줄어들며 속도 또한 저하 될 것이다. 따라서 *부분 형태소 열 사전*에 어떠한 부분 형태소 열과 올바른 형태소 분석 결과를 추가할 것인지 정하는 것은 매우 중요하다.



Target morpheme	Seq. of morphemes	Freq.
한(han)	하(ha)/XSA+ㄴ(n)/ETM	18206
한(han)	하(ha)/XSV+ㄴ(n)/ETM	9865
하고(ha-go)	하(ha)/XSV+고(go)/EC	9494

그림 2. 부분 형태소 열 사전 생성 과정의 예

그림 2는 *부분 형태소 열 사전*을 생성하는 과정을 보여준다. *부분 형태소 열 사전*을 사용하지 않은 좌최장일치법을 이용한 형태소 분석 결과와 학습 말뭉치를 비교하여 오류가 발생한 부분 형태소 열의 정답 부분 형태소 열을 수집하여 사전을 구성한다. 오류가 발생하는 부분 형태소 열을 추출하는 과정에서 발생 빈도를 측정하여 빈도수에 따라 사전의 크기를 정하였다. 그림 3은 부분 결과 사전에 저장할 발생 빈도수의 누적 양을 증가시켰을 때 시스템의 성능 변화를 보여주며 발생 빈도수 누적 양의 상위 40%이상부터 성능 향상의 폭이 현격하게 줄어듦을 확인하였다. 부부 결과 사전은 발생 빈도수의 누적 양이 커질수록 용량이 기하급수적으로 커지기 때문에 사전에 포함될 부분 형태소 열은 40%이하로 제한하였다.

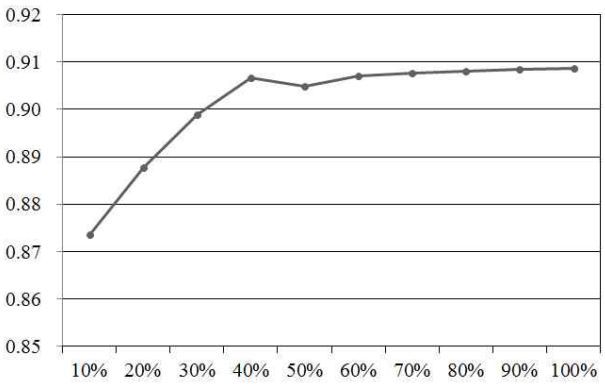


그림 3. 부분 결과 사전 크기에 대한 성능 비교

축소된 HMM 기반 한국어 형태소 품사 부착

제안하는 축소된 HMM 방법은 개선된 좌최장일치법의 결과를 사용하여 형태소 품사 부착을 수행한다. 개선된 좌최장일치법은 기본적으로 1개의 분석 결과를 생성하므로 *부분 형태소 열 사전*에 1개 이상의 후보를 갖는 경우를 제외하면 동일한 형태소 분리를 갖는 결과를 생성한다. 이러한 특징을 이용하여 기존의 HMM을 축소하여 한국어 형태소 품사 부착을 수행한다. 제안한 축소된 HMM 기반 한국어 형태소 품사 부착을 위한 확률 모델은 식 1과 같다.

$$
\begin{aligned}
(\hat{W}, \hat{T}) &= {}^{\mathrm{def}} Argmax_T P(T|W) \\
&= Argmax_T \frac{P(T,W)}{P(W)} \\
&= Argmax_T P(T,W) \\
&\cong Argmax_T \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}), i = o\text{인 경우 } \varnothing
\end{aligned}
\qquad \text{식 1.}
$$

$$
P(w_i|t_i)P(t_i|t_{i-1}) = \begin{cases} P(w_i|t_i)P(t_i|t_{i-1}) \\ < w_i\text{가 한 개의 형태소로 구성되고 여러개의 품사를 갖는 경우} > \\ \prod_{j=1}^{m} P(w_i^j|t_i^j)P(t_i^j|t_i^{j-1}),\ t_i^0 = t_{i-1} \\ < w_i\text{가 } m\text{개의 형태소로 구성된 경우} > \\ C \bullet P(t_i|t_{i-1}) \\ < w_i\text{가 한 개의 형태소와 품사를 갖는 경우} > \end{cases}
$$

HMM은 한국어 형태소 품사 부착에 널리 쓰이는 통계에 기반을 둔 방법이다. HMM을 이용한 형태소 품사 부착은 모든 형태소 분석 결과들 중에서 최적의 형태소 분석 결과를 선택한다(식 1). 식 1에서 W는 주어진 입력 문장의 형태소 순서열, T는 품사 열이다. 식 1은 크게 관측 확률 ($P(w_i/t_i)$)과 전이 확률($P(t_i/t_{i-1})$)로 나눌 수 있다. 이 중 관측 확률은 형태소 사전에 포함된 형태소의 수만큼 확률 값을 저장하여야 하므로 많은 저장 공간을 필요로 한다. 따라서 제안하는 축소된 HMM은 개선된 좌최장일치법과 결합한 제안 모델은 동일한 형태소 분리를 갖는 경우를 고려하여 관측 확률의 저장 공간을 급격하게 줄일 수 있다. 식 1에서 개선된 좌최장

일치법에 의하여 생성된 형태소 분석 결과는 동일한 형태소 분리를 갖기 때문에 입력 문장은 w로 경계가 고정된다. 기존 HMM에서는 주어진 입력 문장에 대하여 최적의 형태소 순서 열과 형태소 품사 열을 찾아야 하지만, 축소된 HMM에서는 고정된 형태소 순서열 w에 대하여 최적의 형태소 품사 열 T를 찾는다. 또한 축소된 HMM은 기존의 HMM과는 다르게 3가지 형태의 확률을 저장한다. 축소된 HMM에서 관측 확률은 형태소 w_i가 여러 품사로 쓰이는 경우 기존의 HMM과 동일한 확률을 가진다. 그리고 형태소 w_i가 여러 개의 형태소로 분리되는 경우는 부분 형태소 열 사전에서 생성하는 후보 들 중에서 결과를 생성한다. 하지만 형태소 w_i가 하나의 품사로만 쓰일 경우는 개선된 좌최장일치법에 의한 형태소 분석 결과가 동일한 분리를 갖기 때문에 하나의 품사로만 쓰이는 형태소의 관측 확률은 최적의 형태소 분석 결과를 찾는데 영향을 미치지 않으며 본 연구에서는 c를 1로 정하였다. 그림 4의 예에서 “압류”는 일반명사만을 품사로 갖기 때문에 “압류/일반명사”의 관측 확률은 최적의 형태소 열을 찾는데 영향을 미치지 않는다. 축소된 HMM은 하나의 품사로만 쓰이는 형태소의 관측 확률을 저장할 필요가 없으며, 하나의 품사로만 쓰이는 형태소의 수는 전체 형태소 수의 약 95%에 해당한다. 따라서 통계 정보를 위해 필요한 저장 공간 및 형태소 품사 부착을 수행 할 때 사용하는 메모리의 양이 급격하게 감소한다는 장점을 가진다.

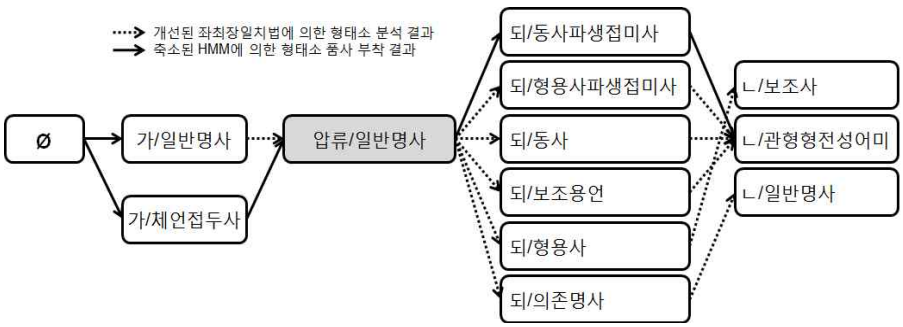


그림 4. “가압류된”의 형태소 품사 부착의 예

실험 및 평가

본 논문에서는 학습 및 실험을 위하여 세종 계획 말뭉치를 사용하였다[17]. 세종 계획 말뭉치는 표 2와 같이 구성되어 있다. 실험에서는 세종 계획 말뭉치 전체를 사용하였고 학습과 검증을 위해 8:2(111,861문장, 1,612,680어절: 27,967문장, 403,180어절)의 비율로 나누어 사용하였다. 실험을 위하여 말뭉치로부터 형태소 접속 규칙 및 형태소 사전을 추출하여 사용하였으며 Uni-gram 관측 확률과 HMM을 위한 형태소 관측 확률 그리고 품사 전이 확률을 maximum likelihood estimator을 사용하여 계산하였다.

표 2. 세종 계획 말뭉치의 구성

분류	개수
문장 수	139,828
어절 수	2,015,860
형태소 수	4,641,546
형태소 태그 수	46

시스템의 성능을 평가하기 위하여 재현율, 정확률 그리고 F_1-평가치를 사용하였으며 표 3은 각 시스템의 성능을 보여준다. 재현율은 시스템이 제안한 형태소, 품사 쌍 중에서 정답 수를 실험 말뭉치에 나타난 형태소의 수로 나눈 것이다. 정확률은 시스템이 제안한 형태소, 품사 쌍 중에서 정답 수를 시스템이 제안한 형태소의 수로 나눈 것이다. 마지막으로 F_1-평가치는 재현율과 정확률의 조화 평균이며 재현율과 정확률의 곱을 재현율과 정확률의 합의 1/2로 계산한 값이다. 표 3은 좌최장일치법을 사용한 시스템, 기존의 HMM 시스템[10] 그리고 제안한 시스템에 대한 재현율, 정확률 그리고 F_1-평가치의 결과를 보여준다.

표 3에서 제안한 시스템은 좌최장일치법을 사용한 시스템보다 F_1-평가치에서 약 8% 높은 성능을 보여주었고 기존 HMM을 이용한 시스템보다는 제안한 시스템이

표 3. 시스템 성능 비교

분류	재현율	정확률	F_1-평가치
좌최장일치법	0.82	0.83	0.82
기존 HMM[10]	0.94	0.94	0.94
제안 방법	0.90	0.92	0.91

표 4. 메모리 사용량과 응답 속도 비교

	좌최장일치법	기존 HMM	제안 방법
어휘사전 (MB)	1.58	1.58	1.58
HMM 통계사전 (MB)	0	1.49	0.27
응답 시간 (sec/sentence)	0.0154	0.1495	0.0195

약 3% 낮은 성능을 보였다. 제안한 시스템은 개선된 좌최장일치법에서 제시한 형태소 분석 결과를 사용하기 때문에 모든 가능한 형태소 분석 결과 중 최적의 형태소 품사 부착 결과를 선택하는 HMM을 이용한 방법보다 근소하게 낮은 성능을 보이지만 제안 시스템은 기존의 HMM을 이용한 방법보다 형태소 분석 및 품사 부착 수행 시간이 월등히 빠르고, 적은 저장 공간을 사용하는 장점이 있다(표 4).

결 론

본 논문에서는 최근 사용이 급속도로 늘고 있는 개인화 소형기기에 적합한 한국어 형태소 분석 및 품사 부착 방법을 제안하였다. 빠른 응답 속도를 위해 좌최장일치법을 응용하여 사용하였고 올바른 형태소 분석 결과가 긴 형태소보다 순위가 낮음으로써 정답으로 생성되지 못하는 단점을 보완하기 위해 *부분 형태소 열 사전*을 이용하였다. 제안하는 방법은 좌최장일치가 하나의 형태소 분석결과만을

제공하는 단점을 해결하면서 수행시간은 큰 차이를 보이지 않았다. 또한 형태소 품사 부착을 위하여 Uni-gram 관측 확률을 이용한 방법과 축소된 HMM을 이용하는 방법을 제안하였다. 축소된 HMM은 기존의 HMM을 변형된 좌최장일치법에 맞게 축소시켜 적은 자원을 사용하면서 HMM의 장점을 유지하는 방법이다. 축소된 HMM은 기존의 HMM을 이용한 시스템과 비교하여 약간의 성능 저하를 보이지만, HMM 통계사전의 크기를 기존 모델의 사전 대비 약 18%로 줄였으며 응답 시간은 약 13%만을 요구하였다.

향후 과제로는 형태소 분석 및 품사 부착 시스템의 자원 사용을 더 줄이기 위하여 형태소 사전의 크기를 줄이는 대신 형태소 사전에 등록되지 않은 미등록어에 대한 형태소 분석 및 품사 부착 방법에 대한 연구가 필요할 것으로 생각한다.

참고문헌

[1] Brill, E. (1995), Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging, *Computational Linguistics, 21-4,* 543-564.

[2] Lovins, J. (1968), Development of a stemming algorithm, *Mechanical Translation and Computational Linguistics, 11,* 22-31.

[3] Lee, E. (1992), *An improved method on Korean morphological analysis based on CYK algorithm*, MS thesis, POSTECH.

[4] Church, K. (1988), A stochastic parts program and noun phrase parser for unrestricted text, *Proc. of Conference on Applied Natural Language Processing*, 136-143.

[5] Kim, J., Lim, H., & Rim, H. (1998), Twoply hidden Markov model: A Korean POS tagging model based on morpheme-unit with Eojeol-unit context, *International Journal of Computer Processing of Oriental Languages, 21-1,* 5-29.

[6] Charniak, E., Hendrickson, C., Jacobson, N., & Perkowitz, M. (1993), Equations for

part of speech tagging, *Proc. of Conference on the American Association for Artificial Intelligence*, 784-789.

[7] Lee, D. & Rim, H. (2005), Probabilistic models for Korean morphological analysis, *Proc. of International Joint Conference on Natural Language Processing*, 197-202.

[8] 이은철, 이종혁 (1992), 계층적 기호 접속정보를 이용한 한국어 형태소 분석기의 구현. **제4회 한글 및 한국어 정보처리 학술 발표 논문집,** 95-104.

[9] 임희석, 김진동, 임해창 (1996), 변형 규칙 기반 한국어 품사 태거의 개선. **제 4회 한글 및 한국어 정보처리 학술 발표 논문집**, 215-221.

[10] 김재훈, 임철수, 서정연 (1995), 은닉 마르코프 모델을 이용한 효율적인 한국어 품사의 태깅. **정보과학회 논문지, 22**(1), 136-146.

[11] Charniac, E. (1996), *Statistical language learning*, MA thesis, MIT press, cambridge.

[12] 김성용 (1996), ***Tabular parsing* 방법과 접속정보를 이용한 한국어 형태소 분석기.** 석사학위논문, 한국과학기술원.

[13] 최재혁, 이상조 (1993), 양방향 최장일치법에 의한 한국어 형태소 분석기에서의 사전 검색 횟수 감소 방안. **정보과학회 논문지, 20**(10), 1497-1507.

[14] Song, Y., Lee, K., & Lee, Y. (1999), Morphological analyzer using longest match method for syntactic analysis, *Proc. of the Morphological Analysis and Tagger Evaluation Contest*, 157-166.

[15] 신준철, 옥철영 (2012), 기분석 부분 어절 사전을 활용한 한국어 형태소 분석기". **정보과학회 논문지: 소프트웨어 및 응용, 39**(5), 415-424.

[16] 신상현, 이근배, 이종혁 (1997), 통계와 규칙에 기반을 둔 2단계 한국어 품사 태깅 시스템. **정보과학회 논문지, 24**(2), 160-169.

[17] Lee, D., & Rim, H. (2005), Probabilistic models for korean morphological analysis, *Companion to the proceedings of the international joint conference on natural language processing*, 197-201.

[18] 이재성 (2011), 한국어 형태소 분석을 위한 3단계 확률 모델. **정보과학회 논문지: 소프트웨어 및 응용, 3**(5), 257-268

[19] 송영길, 김학수 (2009), 한국어 경량형 띄어쓰기 교정 시스템의 구현. **한국컴퓨터교육학회 논문지 12**(2), 87-96.

[20] 김학수 (2012), 저사양 장치에서 자동 띄어쓰기 시스템 구현을 위한 신뢰성 높고 간결한 패턴 매칭 방법. **정보과학회 논문지: 소프트웨어 및 응용 39**(10), 818-823.

[21] The National Institute of the Korean Language (2007), *Final report on achievements of 21st Sejong project: electronic dictionary.*

1 차원고접수 : 2013. 02. 04
2 차원고접수 : 2013. 06. 10
최종게재승인 : 2013. 06. 24

(*Abstract*)

Light Weight Korean Morphological Analysis Using Left-longest-match-preference model and Hidden Markov Model

Sangwoo Kang[1)] Jaechul Yang[2)] Jungyun Seo[1)]

[1)]Computer Science and Engineering Sogang University

[2)]Software Center Samsung Electronics Co.

With the rapid evolution of the personal device environment, the demand for natural language applications is increasing. This paper proposes a morpheme segmentation and part-of-speech tagging model, which provides the first step module of natural language processing for many languages; the model is designed for mobile devices with limited hardware resources. To reduce the number of morpheme candidates in morphological analysis, the proposed model uses a method that adds highly possible morpheme candidates to the original outputs of a conventional left-longest-match- preference method. To reduce the computational cost and memory usage, the proposed model uses a method that simplifies the process of calculating the observation probability of a word consisting of one or more morphemes in a conventional hidden Markov model.

Key words : POS tagging, left-longest-match-preference model, Simplified hidden Markov model, morphological analysis, limited hardware resources